Sketched Ridge Regression:

Kernel and Overdetermined Problems

Alex Gittens
RPI
gittea@rpi.edu

many collaborators
UC Berkeley

SIAM Annual Meeting 2018 | Portland

Kernel Ridge Regression

- Given dataset $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ and kernel function $\kappa(\mathbf{x}_1, \mathbf{x}_2)$, the problem is to solve

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \|\mathbf{K}\boldsymbol{\alpha}-\mathbf{Y}\|_2^2 + \lambda\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}$$

- Optimal solution:

$$\boldsymbol{\alpha}^\star = (\mathbf{K}+\lambda\mathbf{I}_{\mathrm{n}})^{-1}\mathbf{Y}$$

- For large n (i.e. $n\approx 10^6$), $\mathbf{K}$ does not even fit in memory

Iterative Methods

- Since solution doesn't fit in memory, turn to iterative methods
- Classical methods: Conjugate-Gradient, and *Gauss-Siedel*

- We consider randomized block GS (block coordinate descent) for solving positive-definite systems of the form

$$\mathbf{A}\boldsymbol{\alpha}=\mathbf{y}$$

- Given a current iterate

$$(\boldsymbol{\alpha}_{k+1})_{J}=(\boldsymbol{\alpha}_{k})_{J}-\mathbf{A}_{JJ}^{-1}(\mathbf{A}\boldsymbol{\alpha}_{k}-\mathbf{y})_{J}$$

Sampling in Block GS

Two reasonable schemes, given a blocksize p:

- **Fixed Partition**: Divide $[n]$ into blocks $J_1, \ldots, J_{n/p}$ blocks ahead of time. During the iterates, randomly choose a block J_{t_k} where $t_k \sim Unif(\{1,\ldots,n/p\})$.

- **Random coordinates**: At each iteration, choose uniformly from the set $\{J \in 2^{[n]} : |J| = p\}$.

Sampling in Block GS

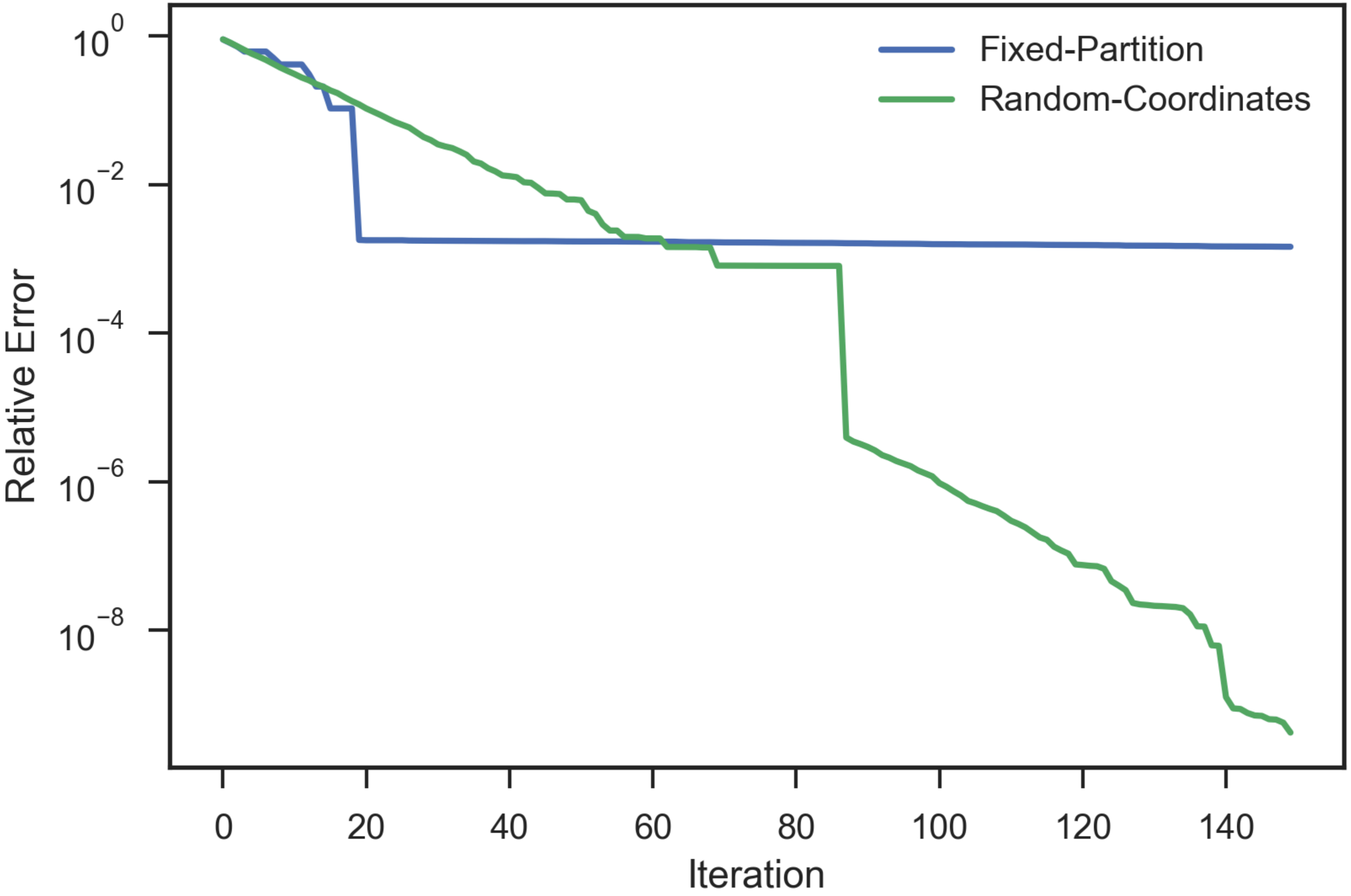
Fixed partitioning is preferable from a systems perspective (cache locality). Random coordinates suffer from slower memory accesses. Why use random coordinates?

A simple example where the sampling makes a large difference: take

$$\mathbf{A}_{\beta} = \mathbf{I} + \beta/n \, \mathbf{1}\mathbf{1}^{T}$$

Try GS with $n=5000$, $p=500$, $\beta=1000$.

Sampling in Block GS



Convergence of Randomized GS

To understand why the behavior differs, look at the theory of randomized GS

Theorem. (Gower and Richtárik, 16)

For all $k \geq 0$,

$$\mathbb{E}\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_*\|_{\mathbf{A}} \leq (1-\mu)^{k/2} \|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_*\|_{\mathbf{A}},$$

where $\mu = \lambda_{min}(\mathbb{E}[\mathbf{P}_{\mathbf{A}^{\mathbf{1/2}}\mathbf{S}}])$. Here, the randomized column selection matrix $\mathbf{S}$ depends on the choice of sampling scheme.

Sampling in Block GS

For our example

$$\mathbf{A}_{\beta} = \mathbf{I} + \frac{\beta}{n}\mathbf{1}\mathbf{1}^{T},$$

$$\mu_{part} = \frac{p}{n+\beta p}$$

$$\mu_{rand} = \mu_{part} + \frac{p-1}{n-1}\frac{\beta p}{n+\beta p}$$

As $\beta \to \infty$, $\mu_{part} \to 1/\beta$ whereas $\mu_{rand} \to p/n$. **This gap is arbitrarily large.**

Sampling Tradeoffs

- **Systems Perspective**: fixed partition sampling is preferable. Can cache blocks ahead of time, replicate across nodes, etc. *Locality is good for performance.*

- **Optimization perspective**: random coordinates is preferable. Each iteration of GS will make more progress. *Locality is bad for optimization.*

What about acceleration?

Add a Nesterov momentum step to the iterates.

- Does the same sampling phenomenon occur with acceleration?
- Does this provide the $\sqrt{\mu}$ behavior we expect?

(Assuming the acceleration parameters are carefully chosen)

Prior State of Theory

The behavior of accelerated **fixed-partition** sampling is understood

Theorem. (Nesterov and Stich, 16)

For all $k \geq 0$, accelerated block GS with fixed-partition sampling satisfies

$$\mathbb{E}\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_*\|_{\mathbf{A}} \lesssim \left(1 - \sqrt{p/n}\,\mu_{part}\right)^{k/2} \|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_*\|_{\mathbf{A}},$$

where $\mu_{part} = \lambda_{min}(\mathbb{E}[\mathbf{P}_{\mathbf{A}^{1/2}\mathbf{S}}])$. Here, the randomized column selection matrix $\mathbf{S}$ corresponds to fixed-partition sampling.

Thus fixed-partition sampling loses a factor of $\sqrt{p/n}$ over the ideal Nesterov rate.

Main Result

Theorem.

For all $k \geq 0$, accelerated block GS with any (non-degenerate) sampling scheme satisfies

$$\mathbb{E}\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_*\|_{\mathbf{A}} \lesssim (1-\tau)^{k/2} \|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_*\|_{\mathbf{A}}.$$

Here $\tau = \sqrt{\mu/\nu}$, where μ is as before and ν is a new quantity which behaves roughly like n/p.

We prove **this rate is sharp**—there exists a starting point which matches the rate up to constants.

Corollaries

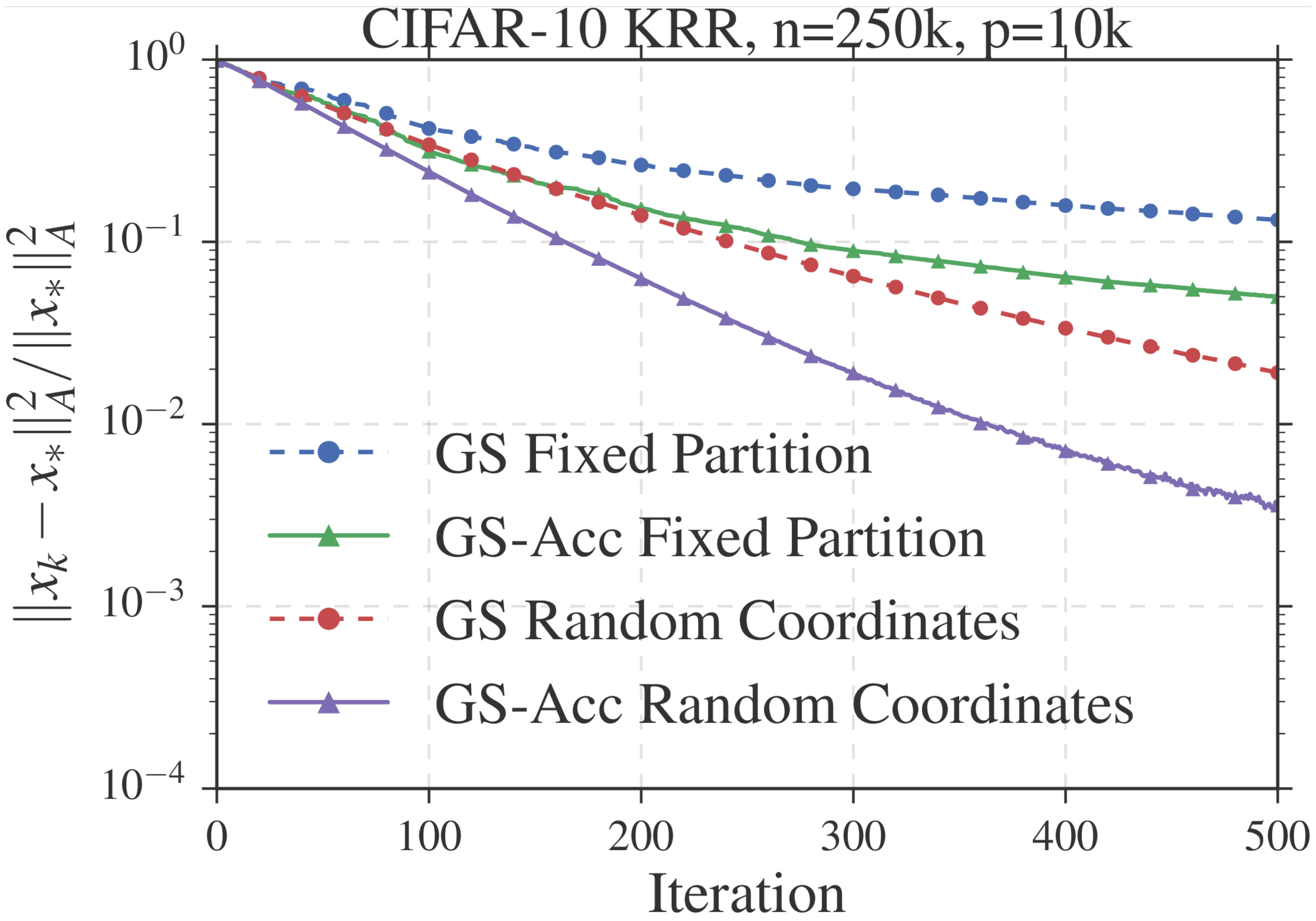
- For fixed partition sampling, we can show that $\nu = n/p$, recovering Nesterov and Stich's earlier result. Combined with the sharpness of the rate, this proves the $\sqrt{p/n}$ loss over the ideal rate is real for the fixed-partition scheme.

- For random coordinate sampling, we can prove the weaker claim

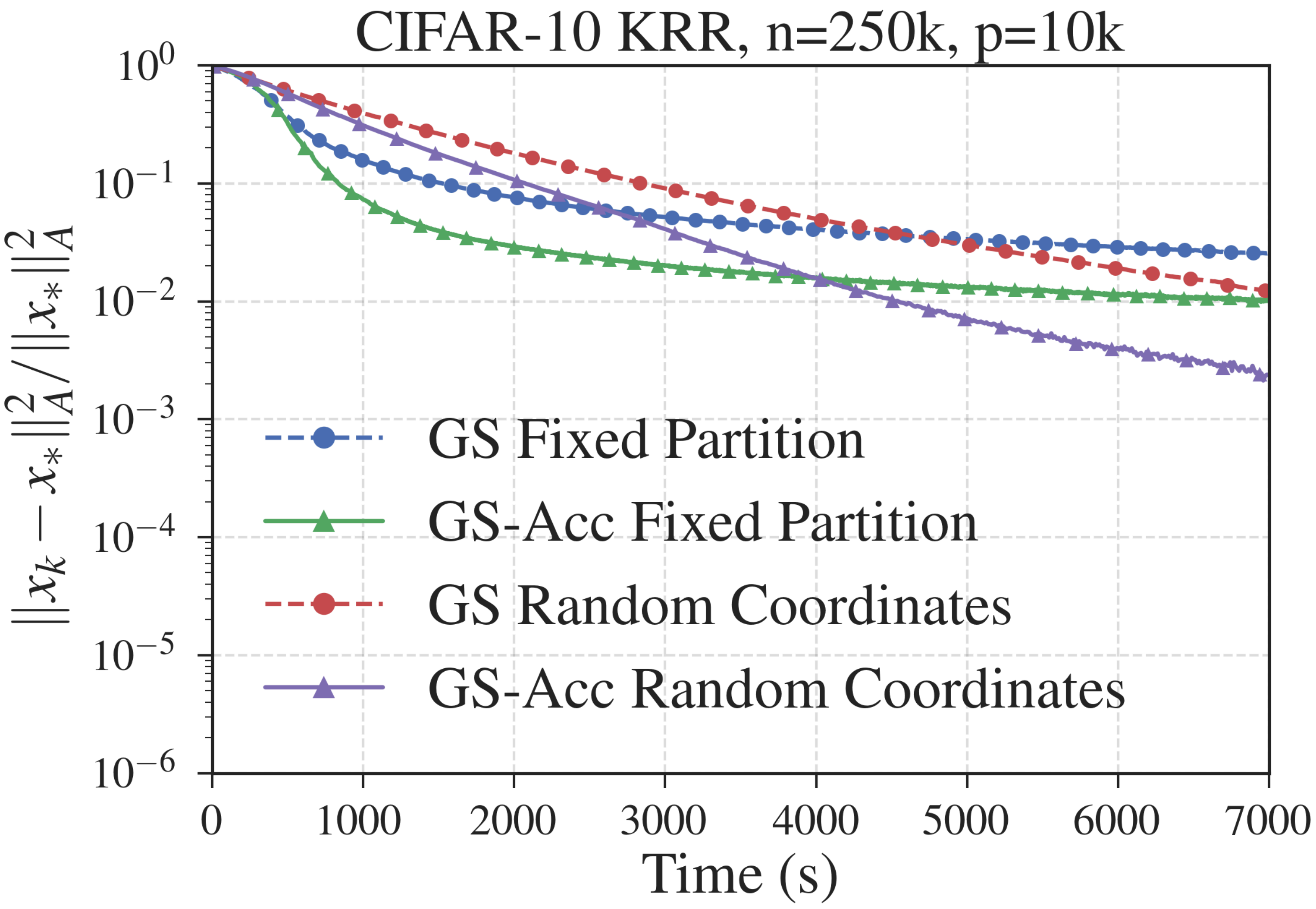
$$\nu \leq \frac{n}{p} \max_{|J|=p} \frac{\max_{i \in J} \mathbf{A}_{ii}}{\lambda_{min}(\mathbf{A}_{JJ})}$$

If all the size J principal submatrices of **A** are sufficiently well-conditioned, $\nu \approx n/p$.

Experiment: Accuracy vs Iteration

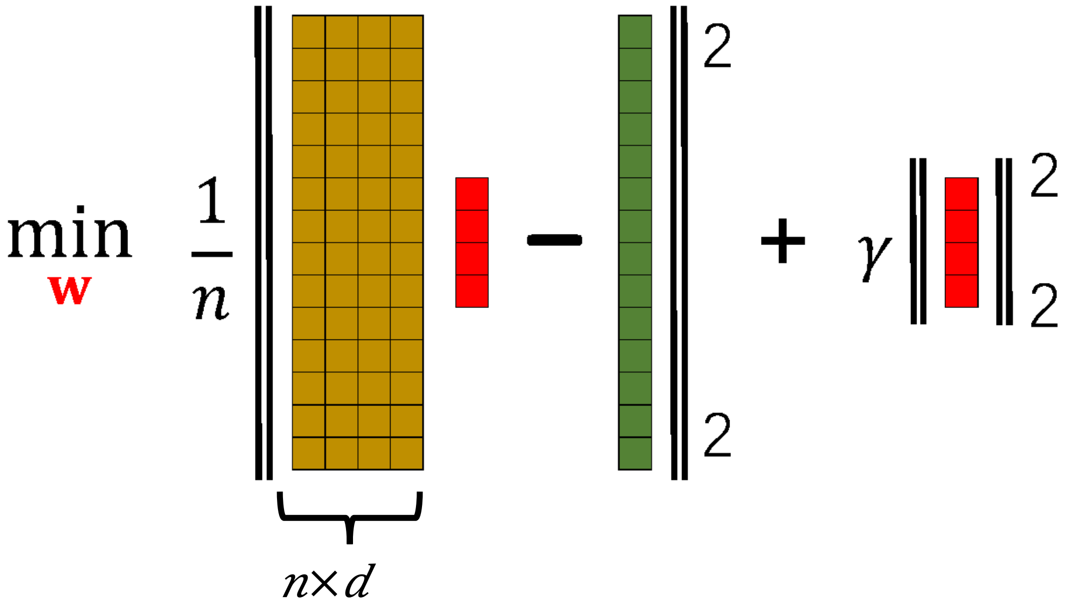


Experiment: Accuracy vs Time



Overdetermined Ridge Regression

$$\min_{\mathbf{w}} \; \{ f(\mathbf{w}) = 1/n \, \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \gamma \|\mathbf{w}\|_2^2 \}$$



Applications:

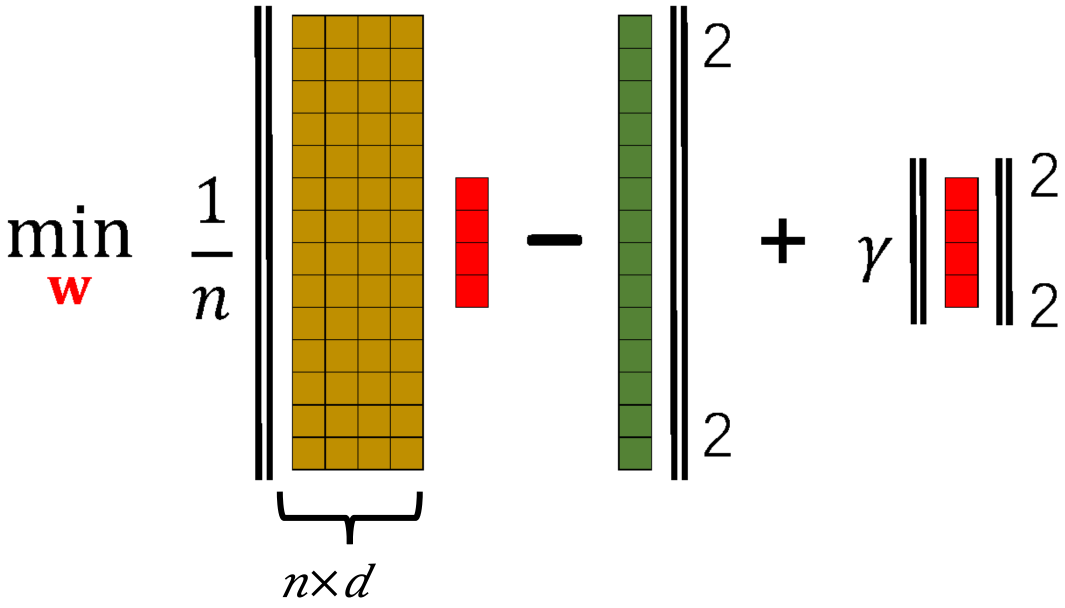
- Basic ML
- IRLS for ℓ_2-penalized GLMs
- Building block in general optimizers

Two Perspectives:

- (Optimization) Deterministic X, y
- (Statistical) Deterministic X, random y

Ridge Regression

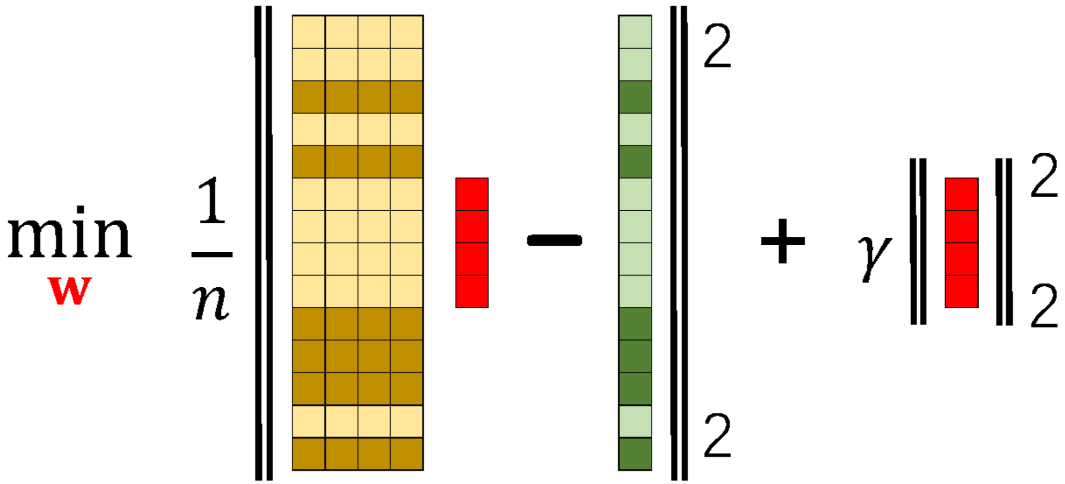
$$\min_{\mathbf{w}} \ \{ f(\mathbf{w}) = 1/n \, \|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \gamma \|\mathbf{w}\|_2^2 \}$$



- Efficient and approximate solution?
- Use only part of the data?

Ridge Regression

min┬**w** {*f*(**w**)=1/*n* ||**Xw**−**y**||↓2↑2 +*γ*||**w**||↓2↑2 }

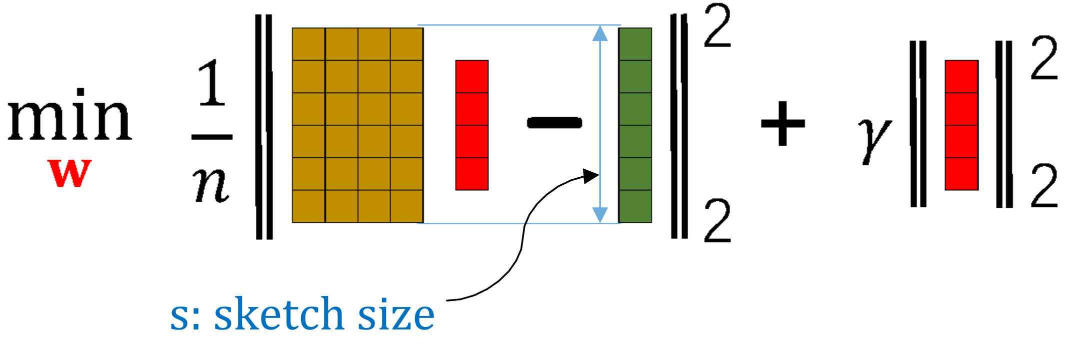


Matrix Sketching:

- Random selection
- Random projection

Approximate Ridge Regression

$\min_{\mathbf{w}} \{ f(\mathbf{w}) = 1/n \, \|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \gamma \|\mathbf{w}\|_2^2 \}$



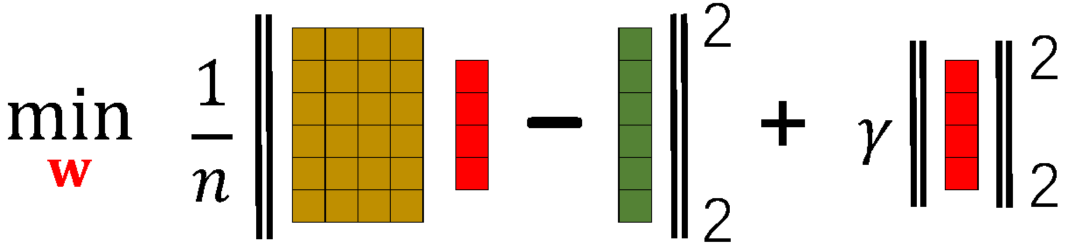
s: sketch size

Optimization Perspective

- Sketched solution: $\mathbf{w}^s$
- $f(\mathbf{w}^s) \le (1+\epsilon) \min_{\mathbf{w}} f(\mathbf{w})$

Approximate Ridge Regression

min┬**w** {*f*(**w**)=1/*n* ||**Xw**−**y**||↓2↑2 +*γ*||**w**||↓2↑2 }



Statistical Perspective

- Bias ||**X***w*↑⋆ −𝔼**Xw**↑s ||↓2
- Variance 𝔼||**X***w*↑s −𝔼**Xw**↑s ||↓2↑2

Related Works on Sketching

Least Squares Regression: $\min_{\mathbf{w}} ||\mathbf{X}\mathbf{w}-\mathbf{y}||_2^2$

Drineas, Mahoney, and Muthukrishnan. *Sampling algorithms for l2 regression and applications*. SODA, 2006.

Clarkson and Woodruff. *Low rank approximation and regression in input sparsity time*. STOC, 2013.

Raskutti and Mahoney. *A statistical perspective on randomized sketching for ordinary least-squares*. JMLR, 2016.

Ridge Regression: $\min_{\mathbf{w}} 1/n \, ||\mathbf{X}\mathbf{w}-\mathbf{y}||_2^2 + \gamma ||\mathbf{w}||_2^2$

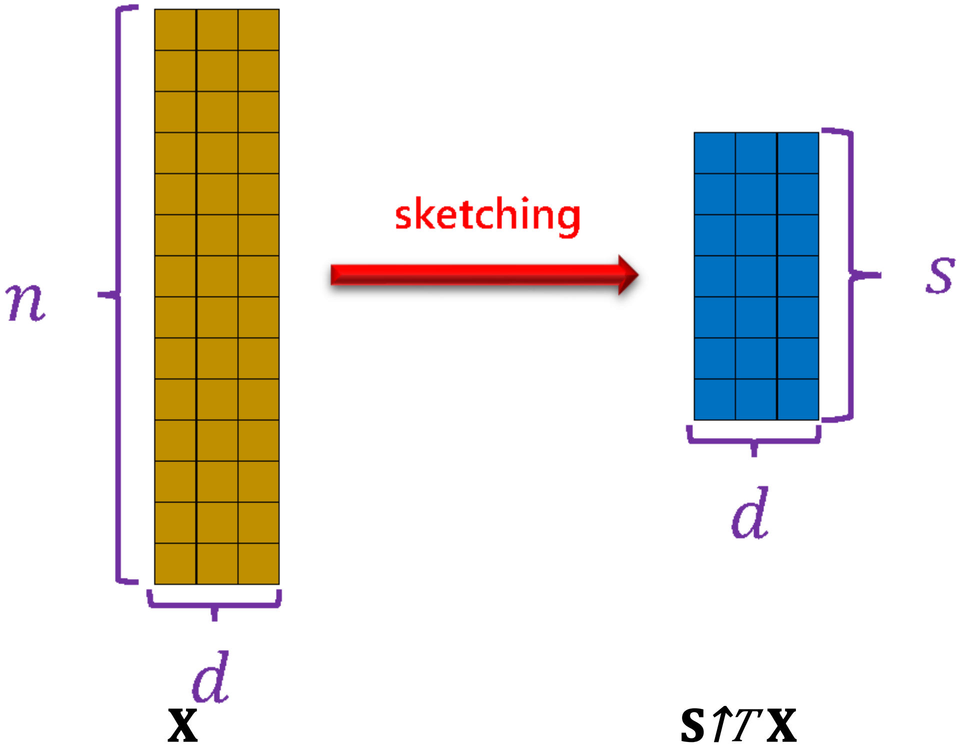
Lu et al. *Faster Ridge Regression via the SRHT*. NIPS, 2013.

Chen et al. *Fast relative-error approximation algorithm for ridge regression*. UAI, 2015.

Avron, Clarkson, Woodruff. *Sharper bounds for Regularized Data Fitting*. Preprint, 2017.

Thanei, Heinze, Meinshausen. *Random projections for large-scale regression*. In Big and Complex Data Analysis, 2017.

Matrix Sketching



- We consider only efficient sketching procedures
  - Time cost is $o(nds)$ — lower than multiplication.
- Examples:
  - Leverage score sampling: $O(nd\log n)$ time
  - SRHT: $O(nd\log s)$ time

Sketched Ridge Regression

- Sketched solution:

$$\mathbf{w}^{s} = \mathrm{argmin}_{\mathbf{w}} \left\{ \frac{1}{n} \|\mathbf{S}^{T}\mathbf{X}\mathbf{w} - \mathbf{S}^{T}\mathbf{y}\|_{2}^{2} + \gamma \|\mathbf{w}\|_{2}^{2} \right\}$$

$$= (\mathbf{X}^{T}\mathbf{S}\mathbf{S}^{T}\mathbf{X} + n\gamma\mathbf{I}_{d})^{\dagger} (\mathbf{X}^{T}\mathbf{S}\mathbf{S}^{T}\mathbf{y})$$

- Time: $O(sd^{2}) + T_{s}$
  - T_{s} is the cost of sketching $\mathbf{S}^{T}\mathbf{X}$
  - E.g. $T_{s} = O(nd \log s)$ for SRHT.
  - E.g. $T_{s} = O(nd \log n)$ for leverage score sampling.
- Versus the time for the full RR problem: $O(nd^{2})$

Results: Optimization Perspective

Optimization Perspective

For the sketching methods

- SRHT or leverage sampling with $s = \mathcal{O}\left(\frac{\beta d}{\epsilon}\right)$,
- uniform sampling with $s = \mathcal{O}\left(\frac{\mu \beta d \log d}{\epsilon}\right)$,

$f(\mathbf{w}^{s}) \leq (1+\epsilon) f(\mathbf{w}^{\star})$ holds w.p. 0.9.

- $\mathbf{X} \in \mathbb{R}^{n \times d}$: the design matrix
- γ: the regularization parameter
- $\beta = \frac{\|\mathbf{X}\|_2^2}{n\gamma + \|\mathbf{X}\|_2^2} \in (0, 1]$
- $\mu \in [1, \frac{n}{d}]$: the row coherence of $\mathbf{X}$

Optimization Perspective

For the sketching methods

- SRHT or leverage sampling with $s = \mathcal{O}(\beta d / \epsilon)$,
- uniform sampling with $s = \mathcal{O}(\mu \beta d \log d / \epsilon)$,

$f(\mathbf{w}^{s}) \leq (1+\epsilon) f(\mathbf{w}^{\star})$ holds w.p. 0.9.

$$\Longrightarrow \quad \frac{1}{n} \|\mathbf{X}\mathbf{w}^{s} - \mathbf{X}\mathbf{w}^{\star}\|_2^2 \leq \epsilon f(\mathbf{w}^{\star}).$$

- $\mathbf{X} \in \mathbb{R}^{n \times d}$: the design matrix
- γ: the regularization parameter
- $\beta = \frac{\|\mathbf{X}\|_2^2}{n\gamma + \|\mathbf{X}\|_2^2} \in (0, 1]$
- $\mu \in [1, n/d]$: the row coherence of $\mathbf{X}$

Results: Statistical Perspective

Statistical Model

- $\mathbf{X} \in \mathbb{R}^{n \times d}$: fixed design matrix
- $\mathbf{w}_0 \in \mathbb{R}^{d}$: the *true* and *unknown* model
- $\mathbf{y} = \mathbf{X}\mathbf{w}_0 + \boldsymbol{\delta}$: observed response vector
  - $\delta_1, \cdots, \delta_n$ are random noise
  - $\mathbb{E}[\boldsymbol{\delta}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\delta}\boldsymbol{\delta}^T] = \xi^2 \mathbf{I}_n$

Bias-Variance Decomposition

- Risk: $R(\mathbf{w}) = 1/n \, \mathbb{E} // \mathbf{Xw} - \mathbf{Xw} \downarrow 0 \; // \downarrow 2 \uparrow 2$
  - $\mathbb{E}$ is taken w.r.t. the random noise $\boldsymbol{\delta}$.

Bias-Variance Decomposition

- Risk: $R(\mathbf{w}) = 1/n\, \mathbb{E} \|\mathbf{X}\mathbf{w} - \mathbf{X}\mathbf{w}_0\|_2^2$
  - $\mathbb{E}$ is taken w.r.t. the random noise $\boldsymbol{\delta}$.
  - Risk measures prediction error.

Bias-Variance Decomposition

- Risk: $R(\mathbf{w}) = 1/n\, \mathbb{E}\|\mathbf{X}\mathbf{w} - \mathbf{X}\mathbf{w}_0\|_2^2$
- $\mathrm{R}(\mathbf{w}) = \mathrm{bias}^2(\mathbf{w}) + \mathrm{var}(\mathbf{w})$

Bias-Variance Decomposition

- Risk: $R(\mathbf{w}) = \frac{1}{n}\mathbb{E}\|\mathbf{X}\mathbf{w} - \mathbf{X}\mathbf{w}_0\|_2^2$
- $\mathrm{R}(\mathbf{w}) = \mathrm{bias}^2(\mathbf{w}) + \mathrm{var}(\mathbf{w})$

Optimal Solution

- $\mathrm{bias}(\mathbf{w}^\star) = \gamma\sqrt{n}\,\|(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_d)^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}_0\|_2$,
- $\mathrm{var}(\mathbf{w}^\star) = \frac{\xi^2}{n}\|(\mathbf{I}_d + n\gamma\mathbf{\Sigma}^{-2})^{-1}\|_2^2$,

Sketched Solution

- $\mathrm{bias}(\mathbf{w}^s) = \gamma\sqrt{n}\,\|(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}\mathbf{\Sigma} + n\gamma\mathbf{I}_d)^\dagger\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}_0\|_2$,
- $\mathrm{var}(\mathbf{w}^s) = \frac{\xi^2}{n}\|(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2})^\dagger\mathbf{U}^T\mathbf{S}\mathbf{S}^T\|_2^2$,

- Here $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the SVD.

Statistical Perspective

For the sketching methods

- SRHT or leverage sampling with $s = \mathcal{O}\left(\frac{d}{\epsilon^2}\right)$;
- uniform sampling with $s = \mathcal{O}\left(\frac{\mu d \log d}{\epsilon^2}\right)$,

- $\mathbf{X} \in \mathbb{R}^{n \times d}$: the design matrix
- $\mu \in \left[1, \frac{n}{d}\right]$: the row coherence of $\mathbf{X}$

the following hold w.p. 0.9:

$$1 - \epsilon \leq \mathrm{bias}(\mathbf{w}^{\mathrm{s}}) / \mathrm{bias}(\mathbf{w}^{\star}) \leq 1 + \epsilon,$$

Good!

$$(1-\epsilon)\frac{n}{s} \leq \mathrm{var}(\mathbf{w}^{\mathrm{s}}) / \mathrm{var}(\mathbf{w}^{\star}) \leq (1+\epsilon)\frac{n}{s}.$$

Bad! Because $n \gg s$.

Statistical Perspective

For the sketching methods

- SRHT or leverage sampling with $s = O(d/\epsilon^2)$,
- uniform sampling with $s = O(\mu d \log d / \epsilon^2)$,

the following hold w.p. 0.9:

$$1-\epsilon \le \mathrm{bias}(\mathbf{w}^{s}) / \mathrm{bias}(\mathbf{w}^{\star}) \le 1+\epsilon,$$

$$(1-\epsilon)n/s \le \mathrm{var}(\mathbf{w}^{s}) / \mathrm{var}(\mathbf{w}^{\star}) \le (1+\epsilon)n/s.$$

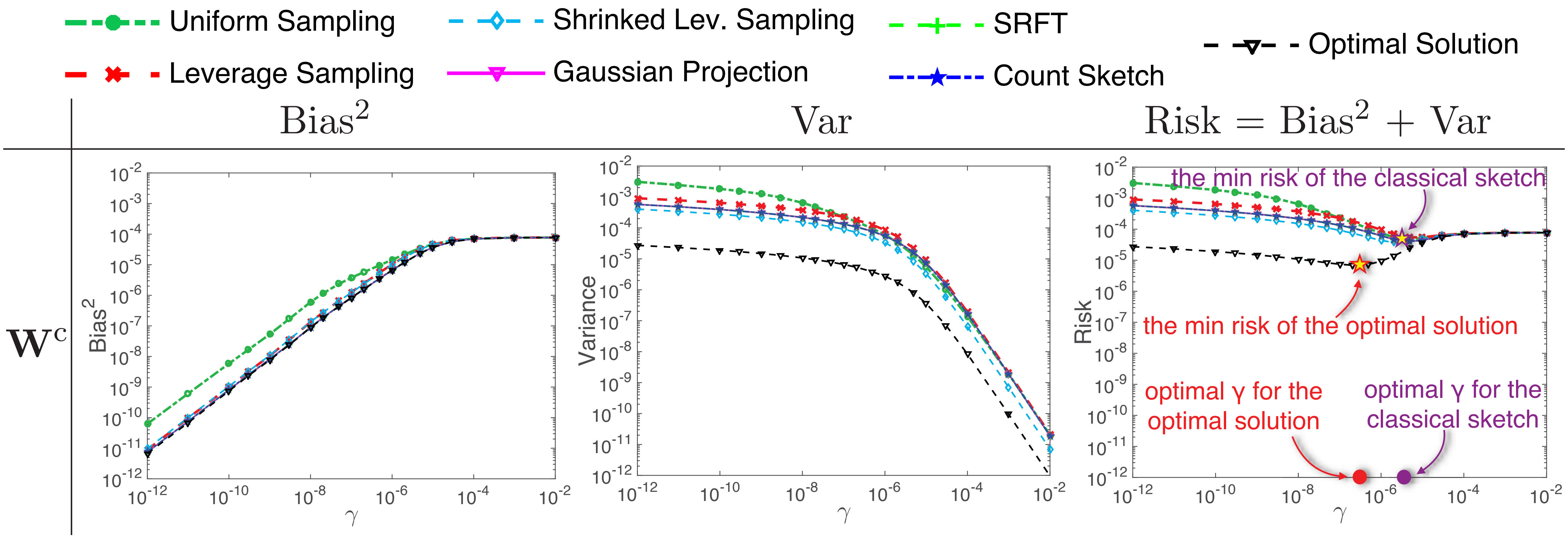
- $\mathbf{X} \in \mathbb{R}^{n \times d}$: the design matrix
- $\mu \in [1, n/d]$: the row coherence of $\mathbf{X}$

If $\mathbf{y}$ is noisy

⇨ variance dominates bias

⇨ $R(\mathbf{w}^{s}) \gg R(\mathbf{w}^{\star})$.

Consequence for selection of regularization



the min risk of the classical sketch

the min risk of the optimal solution

optimal γ for the optimal solution

optimal γ for the classical sketch

Model Averaging to Reduce Variance

Model Averaging

- Independently draw $\mathbf{S}_1, \cdots, \mathbf{S}_g$.
- Compute the sketched solutions $\mathbf{w}_1^{\mathrm{s}}, \cdots, \mathbf{w}_g^{\mathrm{s}}$.
- Model averaging: $\mathbf{w}^{\mathrm{s}} = \frac{1}{g} \sum_{i=1}^{g} \mathbf{w}_i^{\mathrm{s}}$.

Connection to Bagging

- Bagging (bootstrap aggregation) was proposed by Breiman in 1996 for reducing the variance of the decision tree.
- Bagging originates in decision tree methods, but it can be used with many machine learning models.
- For ridge regression, uniform sampling with model averaging is exactly bagging.
- Our approach is not limited to uniform sampling. Random projections and non-uniform sampling outperform uniform sampling.

Optimization Perspective

- For sufficiently large s,

  $f(\mathbf{w}_1^{s}) - f(\mathbf{w}^{\star}) / f(\mathbf{w}^{\star}) \le \epsilon$ holds w.h.p.

Without model averaging

- Using the **same** sketching distribution and s,

  $f(\mathbf{w}^{s}) - f(\mathbf{w}^{\star}) / f(\mathbf{w}^{\star}) \le \epsilon / g + \epsilon^{2}$ holds w.h.p.

With model averaging

Statistical Perspective

- For sufficiently large s, the following hold w.h.p.:

$\mathrm{bias}(\mathbf{w}\uparrow s)/\mathrm{bias}(\mathbf{w}\uparrow\star) \leq 1+\epsilon$ and $\mathrm{var}(\mathbf{w}\uparrow s)/\mathrm{var}(\mathbf{w}\uparrow\star) \leq n/s\ (1+\epsilon)$.

Without model averaging

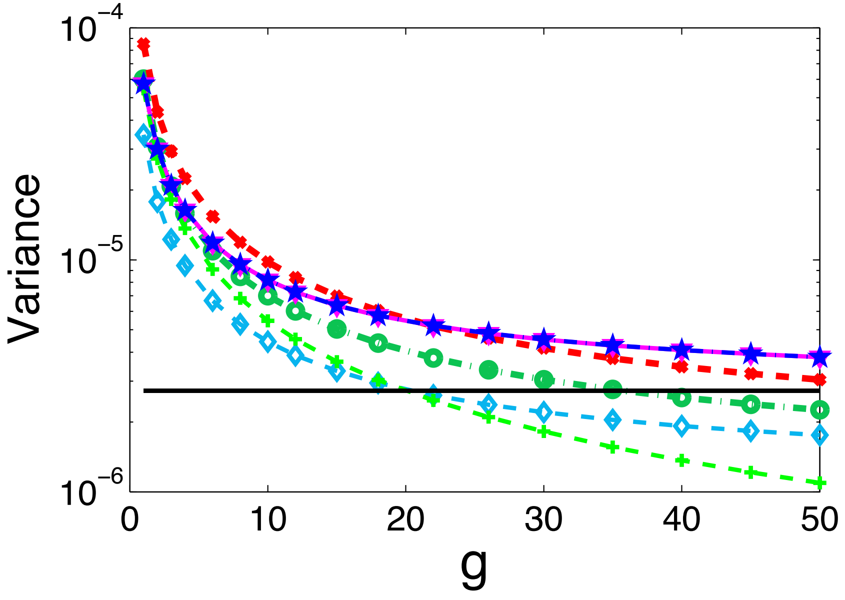
- Using the **same** sketching distribution and s, the following hold w.h.p.:

$\mathrm{bias}(\mathbf{w}\uparrow s)/\mathrm{bias}(\mathbf{w}\uparrow\star) \leq 1+\epsilon$ and $\mathrm{var}(\mathbf{w}\uparrow s)/\mathrm{var}(\mathbf{w}\uparrow\star) \lesssim n/s\ (1/\sqrt{g}+\epsilon)\uparrow\mathbf{2}$

With model averaging

Empirical variance reduction

- If s is large compared to d and g is larger than n/s, then $\mathrm{var}(\mathbf{w}^{\mathrm{s}}) < \mathrm{var}(\mathbf{w}^{\star})$.



Experiments on synthetic data.

- $n = 10^5$, $d = 500$, $\kappa(\boldsymbol{X}^T \boldsymbol{X}) = 10^{12}$.
- Sketch size is $s = 5000 = n/20$.
- Regularization parameter $\gamma = 10^{-6}$.
- As g exceeds $n/s = 20$, $\mathrm{var}(\mathbf{w}^{\mathrm{s}})$ can be smaller than $\mathrm{var}(\mathbf{w}^{\star})$.

Thank You!

Breaking Locality Accelerates Block Gauss-Seidel. Tu, G., et al. ICML 2017
https://arxiv.org/abs/1701.03863

S. Wang, G., and M. W. Mahoney. “*Sketched Ridge Regression: Optimization Perspective, Statistical Perspective, and Model Averaging*”. ICML, 2017.
https://arxiv.org/abs/1702.04837